Acuity: Journal of English Language Pedagogy, Literature, and Culture. Vol.7 No. 1 2022
https://jurnal.unai.edu/index.php/acuity

Sentiment and Sentence Similarity as Predictors of Integrated and Independent L2 Writing Performance

Kutay Uzun,[1]
Trakya University, Department of English Language Teaching, Turkey
kutayuzun@trakya.edu.tr

Ömer Gökhan Ulum[2]
Mersin University, Department of English Language Teaching, Turkey

DOI: 10.35974/acuity.v7i2.2529

Abstract

This study aimed to utilize sentiment and sentence similarity analyses, two Natural Language Processing techniques, to see if and how well they could predict L2 Writing Performance in integrated and independent task conditions. The data sources were an integrated L2 writing corpus of 185 literary analysis essays and an independent L2 writing corpus of 500 argumentative essays, both of which were compiled in higher education contexts. Both essay groups were scored between 0 and 100. Two Python libraries, TextBlob and SpaCy, were used to generate sentiment and sentence similarity data. Using sentiment (polarity and subjectivity) and sentence similarity variables, regression models were built and 95% prediction intervals were compared for integrated and independent corpora. The results showed that integrated L2 writing performance could be predicted by subjectivity and sentence similarity. However, only subjectivity predicted independent L2 writing performance. The prediction interval of subjectivity for independent writing model was found to be narrower than the same interval for integrated writing. The results show that the sentiment and sentence similarity analysis algorithms can be used to generate complementary data to improve more complex multivariate L2 writing performance prediction models.

Keywords: *EFL Writing Performance, Independent Writing, Integrated Writing, Sentiment Analysis, Sentence Similarity, Task Type*

INTRODUCTION

Natural language processing (NLP), which deals with the computational analysis of human languages for both comprehension and production (Crystal, 2008), has been an ever-growing field of research since 1940's. Since then, it has been used for purposes such as machine translation, speech recognition, part-of-speech tagging, sentiment analysis, language production (e.g. chat bots), topic modelling or automated question-answer systems from computer science to political science.

Despite their wide use in various fields, including educational science (e.g. Crossley, Paquette, Dascalu, McNamara & Baker, 2016), foreign language writing research make limited use of state-of-the-art NLP applications in that most studies which utilize NLP seem to benefit from automated feedback/essay evaluation (e.g. Parra & Calero, 2019) and the computation of cohesion (e.g. Jung, Crossley & McNamara, 2019) or complexity indices (e.g. Casal & Lee,

Corresponding Author: Kutay Uzun, Trakya Universitesi, Kosova Yerleskesi, Eğitim Fakultesi, Oda No:G-06.
email: kutayuzun@trakya.edu.tr

2019) with a few exceptions such as DeCoursey and Hamad (2019), Hall and Sheyholislami (2013) and Wang (2020) who investigate sentiments in learner reflections, written feedback and syntactic complexity.

Emotions have been shown to influence second language acquisition (MacIntyre & Gregersen, 2012), vocabulary acquisition (Miller, Fox, Moser & Godfroid, 2018) and performance in foreign language tests and lexical decisions tasks (Dewaele & Alfawzan, 2018). Nonetheless, L2 writing seem to have fallen behind other aspects of language learning in terms of emotion research despite extensive studies on anxiety, a negative emotion, or related constructs such as motivation or attitude. Although these constructs have been studied for decades and fruitful discussions have emerged consequently, it is seen that most of those studies are limited to psychometric scales for the measurement of emotions (e.g. Cheng, 2004; Han & Hiver, 2018); therefore, they are not able to account for the instantaneous variations of those emotions. Moreover, the reflection of emotion or a related construct within the learner text is yet to be discovered except for Wang's (2020) study.

Another problematic area within L2 writing research is cohesion, or the general connectedness of the parts of a text. Traditionally, cohesion is investigated through explicit cues such as conjunctions or personal/demonstrative pronouns. However, cohesion can also be achieved implicitly and this cannot be tracked by traditional means of cohesion assessment. For this reason, certain computationally-available constructs such as type-token ratios, synonym overlap, connective frequency and semantic similarity within (and across, if necessary) texts should be used to assess cohesion (Crossley, Kyle & Dascalu, 2018). However, due to the limited amount of studies regarding each of these constructs, further research is still needed to see how they interact with other constructs regarding L2 writing.

In addition to the necessity to study emotion and cohesion in computational terms, an important distinction in L2 writing lies within the difference between integrated and independent writing tasks, which are inherently different from one another. Integrated writing requires learner-writers to utilize primary and/or secondary sources of information for the completion of the task (Weigle & Parker, 2012). On the contrary, independent writing is exclusively based on the learner-writers personal experiences and available linguistic resources without necessitating any use of sources. As such, it differs from integrated writing in lexical, syntactic and lexicogrammatical terms (Kyle, 2020).

The coverage of academic skills in integrated writing unlike its independent counterpart is among the major differences between two task types (Kyle, 2020). Related to this, integrated writing pieces include more specific lexis, longer words and a lower level of clausal complexity (Cumming et al., 2006; Kyle & Crossley, 2016). Biber, Gray and Staples (2016) also confirm more extensive use of clauses in independent writing and conclude that integrated writing is better marked by nouns, nominals, noun phrases and phrasal complexity. Guo, Crossley and McNamara (2013) also confirm the differences between integrated and independent writing by identifying content word familiarity, content word frequency, third-person singular verbs, base verbs and sentence similarity as predictors of integrated writing scores. On the other hand, independent writing score has been predicted by noun hypernymy, conditional connectives and average syllables per word in their study.

Considering the limited use of NLP technology in foreign language research and the role of emotions in language performance, the amount and scope of the studies dealing with these concepts can be expanded. However, such an expansion should also consider the differences between integrated and independent writing tasks since they bear substantial differences. Therefore, this study aims to contribute to this expansion by searching for the potential connections among L2 writing performance (L2WP), sentiment and sentence similarity as manifested within English as a Foreign Language (EFL) learners' texts, while comparing how these constructs interact with integrated and independent task performance.

Sentiment Analysis and L2 Writing Performance

Sentiment is defined as an individual's emotions, opinions, evaluations or beliefs manifested as language (Wiebe, Wilson, Bruce, Bell & Martin, 2004). Therefore, sentiment analysis (SA) is the systematic analysis of those constructs using NLP methods (Liu, 2010). The analysis of sentiments gives information about the polarity of emotions or opinions as positive, negative or neutral in the form of an index (Munezero, Montero, Sutinen & Pajunen, 2014).

Sentiment analysis typically involves pre-processing and matching or classification stages to produce results. The pre-processing stage involves the removal of stop words (e.g. function words) and symbols and checking the subjectivity of the text. Then, polarity is computed based on a pre-labelled lexicon or machine learning classification algorithms which classify texts using polarity models (Kumar & Teeja, 2012). However, the removal of stop words in the pre-processing stage may not make a significant change in the accuracy of sentiment computation (Jianqiang & Xiaolin, 2017) or even reduce its accuracy (Ghosal, Das & Bhattacharjee, 2015).

Numerous pre-labelled lexicons for sentiment analysis are available in the literature (Liu, 2010). For instance, Linguistic Inquiry and Word Count, The General Inquirer, Hu and Liu's lexicon, The Affective Norms for English Words, SentiWordNet or SenticNet which can also utilize machine learning algorithms such as Naive-Bayes to automate labelling are the widely-used lexicons for sentiment analysis. These lexicons keep large lists of words and their sentiment orientations as classes (e.g. sad: negative, happy: positive) or indices (e.g. great: 3.1, tragedy: -3.4) and sentiment analysis algorithms compare texts to those lists to compute sentiment scores (Hutto & Gilbert, 2014).

Although it is possible to run sentiment analysis with many programming languages, Python-based TextBlob and Vader libraries are the simplest ones to use (Kulkarni & Shivananda, 2019). Both libraries are based on Natural Language Tool Kit (NLTK), which is a high-powered Python package for language processing that is widely-used in research and industry (Bird, Loper & Klein, 2009).

TextBlob produces polarity and subjectivity scores for sentiment analysis. The polarity score is between -1 and 1, -1 indicating total negativity and 1 indicating total positivity. A subjectivity score of 1 indicates total subjectivity while 0 indicates total objectivity (Loria, 2020). A library specifically developed for social media analysis, VADER produces separate positivity, neutrality and negativity scores between 0 and 1. Also, it normalizes these scores into a compound score between -1 and 1, -1 indicating total negativity and 1 indicating total positivity. For analysis, VADER can also use capitalization, punctuation and emoticons (e.g. "This is GOOD!!!" gives a higher positivity score than "This is good.") (Hutto & Gilbert, 2014). Both libraries are widely used in computer science with limited use in other fields such as finance (e.g. Ranjan & Sood, 2019) or education (e.g. Peñafiel, Vásquez, Vásquez, Zaldumbide & Luján-Mora, 2018).

Being related to motivation and self-regulation, emotion is considered as an individual difference in L2 writing (Kormos, 2012). In line with this, most emotion-related L2 writing research focus on anxiety (e.g. Cheng, 2004), attitude (e.g. Yoon & Hirvela, 2004) or motivation (e.g. Lo & Hyland, 2007). Indeed, many studies such as Graham, Berninger and Abbott (2012), Guo (2018) and Graham, Harris, Kiuhara and Fishman (2017) confirm that anxiety, attitude and motivation predict writing performance. Nonetheless, most studies on L2 writing rely solely on psychometric scales to measure emotional constructs; therefore, they cannot track or explain the momentary fluctuations in those emotions, which may affect written production partially or completely. Furthermore, if and how emotions are reflected in the written production itself are mostly left unclear.

Given its potential for computer science, education and even clinical psychology (Provoost, Ruwaard, van Breda, Riper & Bosse, 2019), sentiment analysis can provide information for L2

writing researchers and practitioners regarding how emotions, stances or evaluations are reflected in texts. One such study utilizing SA in L2WP research is that of Wang (2020), which analyses 2620 college-level essays written by Chinese learners of English and reaches the following conclusions:

- Emotions as manifested in texts are influenced by the emotionality of writing topics.
- Textual polarity and syntactic complexity are related.
- Positive and negative emotions cause higher cognitive load and hinder L2WP.
- Optimal performance is achieved through textual neutrality.

To our knowledge, Wang's (2020) study is the only one in the current literature which uses SA in relation to L2WP and it is limited to the syntactic complexity of texts written by Chinese learners of EFL. Findings parallel to Wang's findings in L2 writing can be found in studies which test different skills using non-NLP methods. For instance, the effect of emotions on cognitive load and language performance has been confirmed for L2 listening (Chen & Chang, 2009), reading (Azamnouri, Pishghadam & Meidani, 2020) and vocabulary (Guo, Zou & Peng, 2018). Moreover, lack of objectivity, which is a standard in academic writing (Fulwiler, 2002; Richards & Miller, 2005) and also has cultural roots (Hinkel, 1999; Hwang & Lee, 2008), has been shown to result in lower essay scores among non-native writers of English since it results in an infrequency of proper evidence or justification for claims (Carlson, 1988 as cited in Hinkel, 1999). However, sentiment as measured via sentiment analysis is not a component in these studies and there seems to be no research in the literature regarding the construct and L2WP except for Wang's study, which does not provide comparative results for integrated and independent writing.

Semantic Sentence Similarity and L2 Writing Performance

Semantic similarity is a comparative measure of semantic relatedness which evaluates semantic interactions among language units. In the process, taxonomic relationships and commonality are also considered on a hierarchical basis with corpus-based or knowledge-based methods (Harispe, Ranwez, Janaqi & Montmain, 2015; Turney & Pantel, 2010).

Corpus-based methods extract contextual information from different corpora and use this information to measure semantic relatedness. Knowledge-based methods rely on WordNets, large lexical databases that also keep associations among words, to compute sentence similarity through the hierarchical relations among words. Corpus-based methods are considered more suitable to account for all semantic relations while knowledge-based methods serve better the purpose of encoding hierarchical relations. (Araque, Zhu & Iglesias, 2019). Both methods can be used separately or in combination in word, sentence, paragraph or document levels.

Python libraries such as TextBlob (Loria, 2020), NLTK (Bird et al., 2009) or Spacy (Honnibal & Montani, 2017) can be used for similarity computations with only a few lines of code. These libraries produce scores between 0 and 1 where 0 indicates no similarity and 1 indicates sameness. For instance, the sentences "We should put an end to wars." and "Let's finish wars." produce a similarity score of .87 using Spacy, indicating high similarity. Among NLP libraries, Spacy has been shown to be among the most accurate ones and the fastest one (Honnibal & Johnson, 2015).

Crossley et al. (2018) suggest sentence similarity as an indicator of discourse cohesion. Cohesion refers to the connectedness of texts through surface elements, such as connectives or reference words, which make their meaning more accessible to readers (Bailey, 2011). It is considered to be an integral part of understanding how readers are guided by discourse features towards text comprehension (Baştürkmen & von Randow, 2014). Cohesion can be achieved grammatically through conjunctions, references, substitutions or ellipses, or lexically through collocations and reiterations (Grabe & Kaplan, 2014; Halliday & Matthiessen, 2014). Numerous studies indicate a relationship between cohesion and L2WP (e.g. Crossley et al.,

2018; Crossley, Kyle & McNamara, 2016; McArthur, Jennings & Philippakos, 2019; Yang & Sun, 2012).

Despite the established relationship between cohesion and L2WP, Crossley et al. (2018) warn that the traditional measures of cohesion through overt elements (e.g. use of conjunctions) may be insufficient since it can be achieved explicitly or implicitly (Sanders & Maat, 2006) and in the latter case, the evaluation of cohesion becomes more difficult. For this reason, they propose an NLTK-based tool, TAACO, which assesses local (i.e. sentence-level) and global (i.e. paragraph-level) cohesion through connectives, type-token ratios, lexical overlap and sentence similarity to reveal underlying semantic relations among textual elements which constitute discourse cohesion.

A part of cohesion, sentence similarity has been shown to be related to L2WP. For instance, Crossley and McNamara (2012) reveal a negative correlation between sentence similarity and essay score. In another study, they also find that sentence similarity predicts textual coherence (Crossley & McNamara, 2011). Guo, Crossley and McNamara (2013), Kyle (2020) and Plakans and Gebril (2017) conclude that sentence similarity can predict essay score in integrated tasks. In the light of these findings, sentence similarity is used in automated essay scoring (Roscoe, Crossley, Snow, Varner & McNamara, 2014) and feedback systems (e.g. Lee, Wong, Cheung & Lee, 2009). Nonetheless, Gu et al. (2013) seems to be the only study in the literature that provides comparative results for the predictive strength of sentence similarity in integrated and independent writing. Therefore, more research is thought to be beneficial to understand how sentence similarity interacts with integrated or independent essay quality in different contexts or genres.

Purpose and Research Questions

Considering the absence of a study searching for a link between sentiment and L2WP and the scarcity of those which link sentence similarity and L2WP, this study aims to contribute to the literature by showing if and how sentiment and sentence similarity can predict L2WP while comparing their predictive strengths in integrated and independent writing. The research questions are as follows:

RQ1. Do EFL writers' sentiments as manifested in their essays predict their L2WP?

RQ2. Do the prediction intervals of the sentiment model differ in integrated and independent writing?

RQ3. Do sentence similarity scores of EFL writers predict their L2WP?

RQ4. Do the prediction intervals of the sentence similarity model differ in integrated and independent writing?

METHODS

Due to the computational nature of NLP operations (Crystal, 2008), a quantitative design was preferred. Sentiments, semantic sentence similarities and L2WP were treated numerically.

The Corpora

The corpus of integrated writing samples included 185 literary analysis essays (LAE) previously collected and scored in Author (2019) (n = 125) and Author (in review) (n = 60). It had 61871 words, giving an average of 334.44 words per essay. The essays typically included four to seven paragraphs, responding to an essay question directed towards how a particular theme is handled in a given literary work. As such, the LAE's required writers to make use of primary and secondary sources for completion. The LAE's were scored using the Genre-based Literary Analysis Essay Scoring Rubric (GLAESR). GLAESR is an analytical rubric that is

used to score each rhetorical move in a LAE (stating the background, stating the thesis, presenting arguments, supporting arguments, concluding arguments, consolidating the thesis, stating personal opinion) and produce a total score between 0 and 100 (Author, 2019). In both Author (2019) and Author (in review), scoring demonstrated interrater reliability as confirmed by Spearman's Correlation Coefficients.

For independent writing samples, 500 EFL essays from the International Corpus Network of Asian Learners of English were used (Ishikawa, 2018). The corpus as used in the study consisted of 114996 words with an average of 229.99 words per essay. The essays were reliably scored between 0 and 100 using the ESL Composition Profile which is an analytical rubric that is used to score writing samples according to content, organization, vocabulary, language use and mechanics (Jacobs, Zinkgraf, Wormouth, Hartfiel & Hughey, 1981).

Post-hoc power analysis with G*Power (Faul, Erdfelder, Lang & Buchner, 2007) indicated that the sizes of the corpora were sufficient to achieve 100% statistical power for medium effects in all models.

Both corpora were compiled in higher education contexts.

Data Collection

The data set for the study included the sentiment, sentence similarity and essay scores as provided in the corpora. To avoid computing errors, the authors initially ensured that there was a space after each punctuation mark in the corpus manually and each essay was stored as a .txt file with UTF-8 encoding.

TextBlob was used for sentiment analysis (Loria, 2020); therefore, sentiments were obtained by having an algorithm (APPENDIX A) iterate through all files in the corpus directories and compute the polarity and subjectivity scores for each essay.

For sentence similarity, Spacy was used with its largest model of the English language (en_core_web_lg) (Honnibal & Montani, 2017). To compute a mean sentence similarity value for each essay, an algorithm (APPENDIX B) was written by the authors. The algorithm worked as follows:

1. An essay was read.
2. The sentences in the essay were separated and stored in a list (i.e. tokenization).
3. Each sentence in the essay was compared to all the others in the same essay.
4. The result of each comparison (0.00-1.00) was stored in a list using the following criteria to avoid duplicate comparisons:
    a. Sentence similarity score should have been less than 1.00.
    b. Sentence similarity score (15 digits after decimal point) should not have already been in the list.
5. The mean sentence similarity score was produced for the essay from the sentence similarity scores in the list using NumPy (Oliphant, 2006).
6. The mean sentence similarity score for the essay was stored in a dictionary as "Filename: Sentence similarity Score".
7. The process was repeated for the next essay in the corpus directory.

Data Analysis

The algorithms for the computation of sentiments and sentence similarity were run on Jupyter Notebook (Kluyver et al., 2016). Linear regression analyses were run using JASP v0.12.2 (JASP Team, 2020) to see if sentiment and sentence similarity predicted essay scores since residual distribution in both models were normal (see Table 1), collinearity statistics were not problematic (see Table 2) and there was no heteroscedasticity (Larson-Hall, 2010). For sentiment, a multivariate model which included both polarity and subjectivity as predictor variables were tested. Sentence similarity was tested in a univariate model.

Table 1. Skewness and Kurtosis Values for Model Residuals

Corpus	Model	Skewness	*SE*	Kurtosis	*SE*
Integrated	Sentiment	-0.408	0.179	-0.354	0.355
	Sentence similarity	-0.360	0.179	-0.260	0.355
Independent	Sentiment	-0.489	0.109	1.049	0.218
	Sentence similarity	-0.404	0.109	0.844	0.218

Table 2. Tolerance and Variance Inflation Factor (VIF) Values

Corpus	Model	Variable	Tolerance	*VIF*
Integrated	Sentiment	Polarity	0.967	1.034
		Subjectivity	0.967	1.034
	Sentence similarity	Sentence similarity	N/A	N/A
Independent	Sentiment	Polarity	0.993	1.008
		Subjectivity	0.993	1.008
	Sentence similarity	Sentence similarity	N/A	N/A

Prediction strengths of the models were investigated through their 95% prediction intervals, which provide estimated ranges of actual essay scores with 95% confidence. The difference between the lower and upper bounds in each interval was calculated as the width of the interval, smaller numbers indicating narrower and more precise ranges.

RESULTS

Research Question 1

The first research question aimed to see if sentiment could predict integrated and independent essay scores. The descriptive results are given below in Table 3.

Table 3. Polarity, Subjectivity and Essay Scores

Corpus	Variable	*M*	*SD*	*Min*	*Max*
Integrated	Polarity	0.08	0.15	-0.29	0.43
	Subjectivity	0.51	0.09	0.26	0.74
	Essay Score	55.01	17.98	8.00	97.00
Independent	Polarity	0.12	0.13	-0.31	0.54
	Subjectivity	0.52	0.08	0.29	0.89
	Essay Score	62.77	14.21	7.90	95.00

As seen in Table 3, neither integrated nor independent writing samples were visibly polarized with scores around 0 in both corpora. The subjectivity values in both corpora were also around the midpoint of 0.50.
Regression results for the integrated writing sentiment model are tabulated in Table 4.

Table 4. Regression Results for the Sentiment Model (Integrated)

	SS	*df*	*MS*	*F*	*p*
Regression	2613.257	2	1306.629	4.180	**.017**
Residual	56890.721	182	312.586		
Total	59503.978	184			

R = .210, R^2 = .044, *Adjusted* R^2 = .033, *RMSE* = 17.68

As shown in the table, the multivariate sentiment model which included polarity and subjectivity scores as the predictors of integrated writing essay score was significant, explaining 4.4% of the variance (R^2 = .04, $F_{(2, 182)}$ = 4.18, p < .05). The coefficients for the sentiment model are given below in Table 5.

Table 5. *Coefficients for the Sentiment Model (Integrated)*

Variable	*B*	*SE B*	*β*	*t*	*p*
Constant	74.816	8.141		9.190	**< .001**
Polarity	6.941	8.699	0.059	0.798	.426
Subjectivity	-39.713	15.338	-0.191	-2.589	**.010**

Analyses of the coefficients showed that polarity was not a significant predictor of essay score in the model (t = 0.80, p > .05). However, subjectivity was seen to be a significant negative predictor of integrated essay score (t = -2.59, p = .01).
Regression results for the independent writing sentiment model are tabulated in Table 6.

Table 6. Regression Results for the Sentiment Model (Independent)

	SS	*df*	*MS*	*F*	*p*
Regression	1541.240	2	770.620	3.863	**0.22**
Residual	99153.164	497	199.503		
Total	100694.404	499			

R = .124, R^2 = .015, *Adjusted* R^2 = .011, *RMSE* = 14.12

The regression model showed that the sentiment model could significantly predict independent essay score, explaining 1.5% of the variance (R^2 = .02, $F_{(2, 497)}$ = 3.86, p < .05). The coefficients related to the model are presented below in Table 7.

Table 7. Coefficients for the Sentiment Model (Independent)

Variable	*B*	*SE B*	*β*	*t*	*p*
Constant	40.486	4.035		17.467	**< .001**
Polarity	9.397	5.061	0.083	1.857	.064
Subjectivity	-17.164	7.729	-0.099	-2.221	**.027**

Coefficient analysis showed that polarity was not a significant predictor of independent essay score (t = 1.86, p > .05). On the other hand, Subjectivity was found to be a significant negative predictor of independent essay score (t = -2.22, p < .05).

Research Question 2

The second research question aimed to compare the 95% prediction intervals of the sentiment models for integrated and independent writing. The comparison is tabulated below in Table 8.

Table 8. 95% Prediction Intervals for the Sentiment Model

Essay Score	*M*	*SD*	*Min*	*Max*	*U*	*Z*	*p*	*r*
Integrated	70.33	0.38	69.96	71.94	125250.00	20.113	**< .001**	0.77
Independent	55.67	0.14	55.56	57.18				

As shown in the table, the mean 95% prediction interval for the independent essay scores was 14.66 points narrower than that of the integrated essay scores. The difference was statistically significant with a very large effect (Z = 20.11, p < .001).

Research Question 3

The third research question aimed to see if sentence similarity could predict essay score in integrated and independent writing. The descriptive results are presented below in Table 9.

Table 9. Sentence Similarities and Essay Scores

Corpus	Variable	*M*	*SD*	*Min*	*Max*
Integrated	Sentence Similarity	0.82	0.02	0.79	0.87
	Essay Score	55.01	17.98	8.00	97.00
Independent	Sentence Similarity	0.88	0.01	0.81	0.89
	Essay Score	62.77	14.21	7.90	95.00

Considering that the maximum sentence similarity score could be 1.00, it was seen that the mean sentence similarity score was quite high in the data set for both groups, with a difference of 0.06.
Regression results for the integrated writing sentence similarity model are given below in Table 10.

Table 10. Regression Results for the Sentence Similarity Model (Integrated)

	SS	*df*	*MS*	*F*	*p*
Regression	1482.015	1	1482.015	4.674	**.032**
Residual	58021.964	183	317.060		
Total	59503.978	184			

R = .158, R² = .025, *Adjusted R² =* .020, *RMSE =* 17.81

As seen in the table, sentence similarity could significantly predict essay score in integrated writing, explaining 2.5% of the variance (R^2 = .03, $F_{(2, 182)}$ = 4.67, p < .05). The coefficients for the model are presented in Table 11.

Table 11. Coefficients for the Sentence Similarity Model (Integrated)

Variable	*B*	*SE B*	*β*	*t*	*p*
Constant	-54.531	50.684		-1.076	< **.001**
Sentence Similarity	133.072	61.550	0.158	2.162	**.032**

In the coefficient analysis, it was seen that sentence similarity could predict integrated essay score with a constant of -54.53 and a Beta value of 133.07 ($t = 2.16$, $p < .05$).
The regression results for the independent sentence similarity model are given below in Table 12.

Table 12. *Regression Results for the Sentence Similarity Model (Independent)*

	SS	*df*	*MS*	*F*	*p*
Regression	756.355	1	756.355	3.769	.053
Residual	99938.049	498	200.679		
Total	100694.404	499			

R = .087, R^2 = .008, *Adjusted* R^2 = .006, *RMSE* = 14.17

Analysis revealed that sentence similarity could not significantly predict independent essay score ($F_{(1, 498)} = 3.77$, $p > .05$).

Research Question 4

The fourth research question aimed to compare the 95% prediction intervals related to the sentence similarity models of integrated and independent writing. However, no comparison could be made since the variable could not significantly predict independent essay score. The 95% prediction intervals for the integrated essay scores in the data set were found to have a mean of 70.64 (*SD* = 0.14) with a minimum of 70.45 and a maximum of 71.18 points.

DISCUSSION

The study aimed to find out if sentiment and sentence similarity, computed via NLP methods, could predict integrated and independent L2WP. The results showed that the polarity component of sentiment could not predict L2WP in either task type; however, subjectivity was a significant negative predictor of both integrated and independent L2WP with a very small effect. The comparison of 95% prediction intervals showed that subjectivity as a negative predictor could predict L2WP in a much narrower range in independent writing.
The second major finding obtained in the study was that mean sentence similarity could predict integrated L2WP significantly with a very small effect. The variable could not predict independent L2WP.
The differences in integrated and independent writing as observed in the analyses confirmed Biber et al. (2016), Cumming et al. (2006), Kyle (2020) and Kyle and Crossley (2016) who also indicated varying features of the two task/L2WP types. Apparently, learner-writers undergo different thinking and written production processes during integrated and independent writing and this results in visible differences in terms of language use manifested as certain constructs such as word familiarity, verb use, subjectivity and sentence similarity.
Regarding sentiment, it is known that emotions, stances or personal evaluations are among the individual differences in L2 writing (Kormos, 2012) and these constructs seem to be reflected in texts written by learners, making a difference in their L2WP. In the present study,

subjectivity was found to be a negative predictor of both integrated and independent writing, signalling that more subjective essays received lower scores. This finding can be considered parallel to that of Wang (2020) although it is limited to syntactic complexity. In both studies, and regardless of task type in the present study, textual objectivity seemed to result in increased performance.

The reason why higher objectivity results in better performance in both integrated and independent writing can be related to the objectivity standard in essay writing (Fulwiler, 2002; Richards & Miller, 2005) as well as an increased cognitive load due to the emotionality as observed in learner texts. As suggested by Carlson (1988) and Hinkel (1999), a lack of objectivity in writing may indicate weaknesses in crucial concepts such as evidence or justification in texts. Considering that both integrated and independent corpora consisted of expository/argumentative writing tasks, evidence and justification was a required component in all essays. Successful justification of claims with or without source texts naturally requires an objective outlook which would allow learner-writers to present their arguments from multiple perspectives. In that respect, a high level of subjectivity may be signalling a lack of these justifications, resulting in lower essay scores in both integrated and independent writing. Moreover, positive and negative emotions increase cognitive load as concluded by Wang (2020). Defined in relation to working memory (Cooper, 1998), cognitive load is a crucial factor in L2WP because L2 writing, by itself, can overload working memory due to the intensity of the mental processes involved, resulting in poor performance and frequent errors (Nawal, 2018). In addition to the natural cognitive load of L2 writing, the added load due to the emotionality manifested as subjectivity in texts may have further impeded the working memory, resulting in lower scores in both corpora.

Subjectivity as a negative predictor demonstrated higher prediction precision in independent writing than integrated writing. Although the data set used in this study is not sufficient to explore the reasons behind this difference, a plausible explanation may be that the source-based requirements of the literary analysis essay more readily push learners towards a certain level of objectivity while independent writing may be more flexible in that regard, allowing the learner-writer approach the objectivity issue more liberally while writing an essay based on life experiences and opinions. This may, therefore, result in a larger negative effect of subjectivity on essay scores since its excess has been documented to result in lower scores in early studies as well (e.g. Carlson, 1988). However, I believe a cross-comparison of integrated and independent writing samples in terms of objectivity and lexicogrammatical features is necessary for a more assertive conclusion.

The results revealed sentence similarity as a positive predictor of integrated L2WP. However, the construct was not a significant predictor of independent L2WP. This finding corroborated those of Guo et al. (2013) which indicated the same result. In their study, Guo et al. explain the differences through the life experience and personal opinion-based nature of independent writing and the source-based nature of integrated writing which allows learner-writers to use the sources as models. Moreover, sentence similarity is a component within textual cohesion. Considering this, the necessity to integrate sources to produce a whole in integrated writing may be pushing writers to write more cohesive essays, which is also the case in expository writing (Crossley, 2013; Guo et al, 2013). Considering that the integrated writing corpora used in this study consisted exclusively of expository literary analysis essays, the same reason may have applied for the finding that revealed sentence similarity as a significant predictor of integrated writing performance but not of independent writing performance. As such this finding was also in line with Kyle (2020), Plakans and Gebril (2017) and Crossley and McNamara (2012), the last one of which indicating no relationship between sentence similarity and independent writing performance.

CONCLUSION

The results of the study show that the subjectivity component of sentiment analysis can predict both integrated and independent L2 writing performance. In both task conditions, subjectivity serves as a negative predictor, indicating that more objective texts receive higher scores. The results also indicate the sentence similarity predicts only integrated L2 writing performance while it does not seem to be related to independent writing. As such, the findings bear importance as to the use of sentiment analysis in L2 writing performance research and confirm the previously proposed use of sentence similarity analysis within the same domain.

Bearing the findings in mind, consciousness-raising interventions can be developed and applied by teachers and researchers to improve objectivity and integratedness in learner writing. Although effect sizes of the prediction equations were quite small in this study, the results revealed the contribution of these constructs to L2WP.

The small effect sizes of the regression models should be treated with caution since prediction intervals in all models were rather wide in both integrated and independent corpora. In this regard, it is not recommended to attempt score predictions based solely on these variables. Instead, the variables should be seen complementary to more complex multivariate prediction models.

Apart from sentiment and sentence similarity in particular, the results also confirm NLP in general as a beneficial tool for researchers of language learning/teaching as well as practitioners. Using NLP tools for the analysis of learner language seems to provide insights that may not be accessible through more traditional forms of data collection. Both automated and manual forms of written corrective feedback or assessment can benefit from the indices produced thanks to these tools.

As shown in the literature and this study, task type influences how different variables interact with L2WP. In that respect, different genres should be tested using the same methodology for comparison purposes. Moreover, the data set used in this study cannot explain why objectivity can produce a narrower prediction interval for independent writing than integrated writing. For a thorough explanation, the lexicogrammatical features of highly objective and highly subjective texts should be compared in integrated and independent task conditions.

REFERENCES

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems, 165*, 346-359.

Azamnouri, N., Pishghadam, R., & Meidani, E. N. (2020). The role of emotioncy in cognitive load and sentence comprehension of language learners. *Issues in Language Teaching, 9*(1), 29-55. https://doi.org/10.22054/ilt.2020.51543.485

Bailey, S. (2011). *Academic writing: A handbook for international students* (3rd ed.). Abingdon/New York, NY: Routledge.

Baştürkmen, H., & von Randow, J. (2014). Guiding the reader (or not) to re-create coherence: Observations on postgraduate student writing in an academic argumentative writing task. *Journal of English for Academic Purposes, 16*, 14-22.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*(5), 639-668.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python*. Sebastopol: O'Reilly Media Inc.

Carlson, S. (1988). Cultural differences in writing and reasoning skills. In A. C. Purver (Ed), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 109-137). Newbury Park, CA: Sage.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing, 44*, 51-62.

Chen, I., Chang, C. (2009). Cognitive Load Theory: An Empirical Study of Anxiety and Task Performance in Language Learning. *Electronic Journal of Research in Educational Psychology, 7*(18), 729-746. http://dx.doi.org/10.25115/ejrep.v7i18.1369

Cheng, Y. S. (2004). A measure of second language writing anxiety: Scale development and preliminary validation. *Journal of Second Language Writing, 13*(4), 313-335. https://doi.org/10.1016/j.jslw.2004.07.001

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second language writing*. Cambridge, England: CUP.

Cooper, G. (1998, December). *Research into cognitive load theory and instructional design at UNSW*. Sydney, Australia: University of New South Wales. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.3428&rep=rep1&type=pdf

Crossley, S. & McNamara, D. (2011). Text coherence and judgments of essay quality: models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46*(2), 256-271.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-135.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 6-14). ACM. https://doi.org/10.1145/2883851.2883931

Crystal, D. (2008) *Dictionary of linguistics and phonetics* (6th ed.). Oxford: Blackwell.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Research Report Series, 2005*(1), i-77.

DeCoursey, C. A., & Hamad, A. N. (2019). Emotions across the essay: What second-language writers feel across four weeks' writing a research essay. *English Studies at NBU, 5*(1), 114-134. https://doi.org/10.33919/esnbu.19.1.6

Dewaele, J. M., & Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance? *Studies in Second Language Learning and Teaching, 8*(1). https://doi.org/10.14746/ssllt.2018.8.1.2

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fulwiler, T. (2002). *College writing: A personal approach to academic writing* (3rd ed.). Portsmouth, NH: Heinemann Boynton/Cook

Ghosal, T., Das, S. K., & Bhattacharjee, S. (2015). Sentiment analysis on (Bengali horoscope) corpus. In *2015 Annual IEEE India Conference (INDICON)* (pp. 1-6), New Delhi, India. https://doi.org/10.1109/INDICON.2015.7443551.

Grabe, W., & Kaplan, R. B. (2014). *Theory and Practice of Writing*. Abingdon/New York, NY: Routledge.

Graham, S., Berninger, V., & Abbott, R. (2012). Are attitudes toward writing and reading separable constructs? A study with primary grade children. *Reading & Writing Quarterly, 28*(1), 51–69.

Graham, S., Harris, K. R., Kiuhara, S. A., & Fishman, E. J. (2017). The relationship among strategic writing behavior, writing motivation, and writing performance with young, developing writers. *The Elementary School Journal, 118*(1), 82-104.

Guo, J. D. (2018). Effect of EFL writing self-concept and self-efficacy on writing performance: Mediating role of writing anxiety. *Foreign Language Research, 2*, 69-74.

Guo, J., Zou, T., & Peng, D. (2018). Dynamic influence of emotional states on novel word learning. *Frontiers in Psychology, 9*, 1-12. https://10.3389/fpsyg.2018.00537

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218–238. https://10.1016/j.asw.2013.05.002

Hall, C., & Sheyholislami, J. (2013). Using appraisal theory to understand rater values: An examination of rater comments on ESL test essays. *Journal of Writing Assessment, 6*(1), 1-17.

Halliday, M. A., & Matthiessen, C. M. (2014). *Halliday's introduction to functional grammar*. Oxford: Routledge.

Han, J., & Hiver, P. (2018). Genre-based L2 writing instruction and writing-specific psychological factors: The dynamics of change. *Journal of Second Language Writing, 40*(1), 44-59. https://doi.org/10.1016/j.jslw.2018.03.001

Harispe S., Ranwez S., Janaqi S., & Montmain J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies, 8*, 1–254. https://doi.org/10.2200/S00639ED1V01Y201504HLT027

Hinkel, E. (1999). *Objectivity and credibility in L1 and L2 academic writing. Culture in second language teaching and learning*. Cambridge: Cambridge University Press.

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In L. Màrquez, C. Callison-Burch, & J. Su (eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373-1378). Lisbon, Portugal: Association for Computational Linguistics.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI

Hwang, S. & Lee, M. (2008). Syntactic and referential markers ensuring objectivity in EFL essay writing. *English Teaching, 63*(4), 29-47.

Ishikawa, S. (2018). The ICNALE edited essays; A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Studies, 25*, 117-130.

Jacobs, H.L., Zinkgraf, S.A., Wormouth, D.R., Hartfiel, V.F., & Hughey, J.B. (1981). *Testing ESL composition: A practical approach*. Rowely, MA: Newbury House.

JASP Team (2020). *JASP* (Version 0.12.2)[Computer software]. Retrieved from https://jasp-stats.org/

Jianqiang, Z., & Xiaolin, G. (2017). Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*, *5*, 2870-2879. https://doi.org/10.1109/ACCESS.2017.2672677.

Jung, Y., Crossley, S., & McNamara, D. (2019). Predicting Second Language Writing Proficiency in Learner Texts Using Computational Tools. *The Journal of Asia TEFL, 16*(1), 37-52. https://dx.doi.org/10.18823/asiatefl.2019.16.1.3.37

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Ivanov, P. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt, Positioning and power in academic publishing: Players, agents and agendas (pp. 87-90). Amsterdam: IOS Press.

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing, 21*, 390-403. https://doi.org/10.1016/j.jslw.2012.09.003.

Kulkarni, A., & Shivananda, A. (2019). *Natural Language Processing Recipes. Unlocking Text Data with Machine Learning and Deep Learning using Python*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-4267-4.

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis: A perspective on its past, present and future. *International Journal of Intelligent Systems and Applications, 4*(10), 1-14.

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, 100467. https://doi.org/10.1016/j.asw.2020.100467

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12-24.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge

Lee, C., Wong, K. C., Cheung, W. K., & Lee, F. S. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning, 22*(1), 57-72.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (eds.), *Handbook of Natural Language Processing* (2nd ed.) (pp. 627–666). Boca Raton, FL: Chapman & Hall/CRC.

Lo, J., & Hyland, F. (2007). Enhancing students' engagement and motivation in writing: The case of primary students in Hong Kong. *Journal of Second Language Writing, 16*, 219-237. https://doi.org/10.1016/j.jslw.2007.06.002

Loria, S. (2020). *TextBlob Documentation* (Release 0.16.0). Retrieved from https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf

MacIntyre, P. D., & Gregersen, T. (2012a). Emotions that facilitate language learning: The positive-broadening power of the imagination. *Studies in Second Language Learning and Teaching, 2*, 193-213. doi: 10.14746/ssllt.2012.2.2.4

McArthur, C.A., Jennings, A. & Philippakos, Z.A. (2019) Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing, 32*, 1553–1574. https://doi.org/10.1007/s11145-018-9853-6

Miller, Z. F., Fox, J., Moser, J. S., & Godfroid, A. (2018). Playing with fire: Effects of negative mood induction and working memory on vocabulary acquisition. *Cognition and Emotion, 32*, 1105–1113. https://doi.org/10.1080/02699931.2017.1362374.

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions of Affective Computing, 5*(2), 101-111. https://doi.org/10.1109/TAFFC.2014.2317187.

Nawal, A. F. (2018). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies, 3*(1), 385-402. https://doi.org/10.1080/23752696.2018.1513812

Oliphant, T. E. (2006). *A guide to NumPy*. USA: Trelgol Publishing.

Parra G., L., & Calero S., X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction, 12*(2), 209-226. https://doi.org/10.29333/iji.2019.12214a

Peñafiel, M., Vásquez, S., Vásquez, D., Zaldumbide J., & Luján-Mora, S. (2018). Data Mining and Opinion Mining: A Tool in Educational Context. In *Proceedings of the 2018 International Conference on Mathematics and Statistics (ICoMS 2018)* (pp. 74-78). New York, NY: Association for Computing Machinery https://doi.org/10.1145/3274250.3274263

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*, 98–112. doi:10.1016/j.asw.2016.08.005

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study [Provisional PDF]. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01065

Ranjan, S., & Sood, S. (2019). Investor community sentiment analysis for predicting stock price trends. *International Journal of Management, Technology and Engineering, 9*(5), 6012-6020.

Richards, J. C., & Miller, S. K. (2005). *Doing academic writing in education: Connecting the personal and the professional.* Mahwah, NJ: Erlbaum

Roscoe, R. D., Crossley, S. A., Snow, E. L., Varner, L. K., & McNamara, D. S. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. In W. Eberle & C. Boonthum-Denecke (eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 393-398). Palo Alto, CA: AAAI Press.

Sanders, T., & Maat, H. P. (2006). Cohesion and coherence: Linguistic approaches. *Reading, 99*, 440–466.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Wang, Y. (2020). Emotion and syntactic complexity in L2 writing: A corpus-based study on Chinese college-level students' English writing. *The Asian Journal of Applied Linguistics, 7*(1), 1-17.

Weigle, C. S., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*(2), 118-133. https://doi.org/10.1016/j.jslw.2012.03.004.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational linguistics, 30*(3), 277-308.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and education, 23*(1), 31-48.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing, 13*, 257–283. https://doi.org/10.1016/j.jslw.2004.06.002

APPENDIX A. Sentiment Analysis Algorithm

```
import os
import glob
from textblob import TextBlob
#Researchers can use the same algorithm by simply changing the file path below.
os.chdir(r'C:\Corpus_Directory')
corpus = glob.glob('*.txt')
for essay in range(len(corpus)):
    f = open(corpus[essay], encoding='utf-8')
    content = f.read()
    text = TextBlob(content)
    sentiment_score = text.sentiment
    f.close()
    print(corpus[essay], sentiment_score)
```

APPENDIX B. Sentence Similarity Algorithm

```
import os
import glob
import spacy
import numpy as np
nlp = spacy.load("en_core_web_lg")
#Researchers can use the same algorithm by simply changing the file path below.
os.chdir(r'C:\Corpus_Directory')
corpus = glob.glob('*.txt')
similarity_list = []
similarity_results = {}
for essay in range(len(corpus)):
    f = open(corpus[essay], encoding='utf-8')
    content = f.read()
    doc = nlp(content)
    sentences = list(doc.sents)
    for sentence1 in sentences:
        for sentence2 in sentences:
            similarity = sentence1.similarity(sentence2)
            if similarity < 1.0 and similarity not in similarity_list:
                similarity_list.append(similarity)
                similarity_results[f] = np.mean(similarity_list, dtype=np.float64)
                f.close()
```